

Secure Erasure Codes With Partial Decodability

Son Hoang Dau*, Wentu Song[†], Chau Yuen[‡]

Singapore University of Technology and Design, Singapore

Emails: {^{*}sonhoang_dau, [†]wentu_song, [‡]yuenchau}@sutd.edu.sg

Abstract—The MDS property (aka the k -out-of- n property) requires that if a file is split into several symbols and subsequently encoded into n coded symbols, each being stored in one storage node of a distributed storage system (DSS), then an user can recover the file by accessing any k nodes. We study the so-called p -decodable μ -secure erasure coding scheme ($1 \leq p \leq k - \mu, 0 \leq \mu < k, p|(k - \mu)$), which satisfies the MDS property and the following additional properties:

- (P1) **strongly secure up to a threshold**: an adversary which eavesdrops at most μ storage nodes gains no information (in Shannon's sense) about the stored file,
- (P2) **partially decodable**: a legitimate user can recover a subset of p file symbols by accessing some $\mu + p$ storage nodes.

The scheme is *perfectly* p -decodable μ -secure if it satisfies the following additional property:

- (P3) **weakly secure up to a threshold**: an adversary which eavesdrops more than μ but less than $\mu + p$ storage nodes cannot reconstruct any part of the file.

Most of the related work in the literature only focused on the case $p = k - \mu$. In other words, no partial decodability is provided: an user cannot retrieve any part of the file by accessing less than k nodes. For our more general code, Property (P2) guarantees partial decodability: once the user accesses p more nodes than the strong security threshold μ , it can start to decode some subset of p file symbols.

We provide an explicit construction of p -decodable μ -secure coding schemes over small fields for all μ and p . That construction also produces *perfectly* p -decodable μ -secure schemes over small fields when $p = 1$ (for every μ), and when $\mu = 0, 1$ (for every p). We establish that perfect schemes exist over *sufficiently large* fields for almost all μ and p .

I. INTRODUCTION

Data replication is the most common way for distributed storage systems (DSS) to guarantee high data availability and node failure tolerance. Most of the current distributed storage systems are using 3-way replication where each piece of data is replicated three times and each of its copy is stored at a different storage node in the system. If at most two storage nodes are down, the data is still available at at least one node. However, 3-way replication is highly inefficient in storage overhead, as only a very modest portion 33% of the whole storage capacity can be used. As the demand for data storage scales up quickly, replication based storage systems incur significantly high cost in terms of storage footprint and power usage for cooling systems. It is well known that erasure codes [1] possess lots of advantages over replication [2]. Giants such as Microsoft, Facebook, and Google have, therefore, included erasure codes, alongside replication, in their distributed storage systems [3], [4], [5].

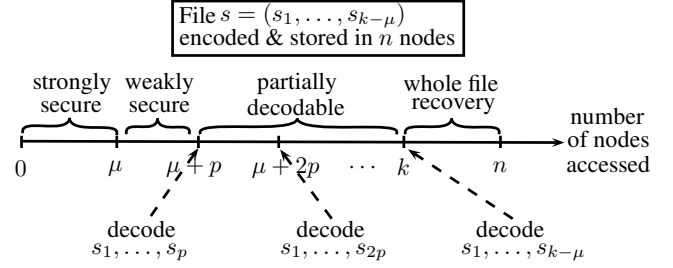


Fig. 1: Illustration for perfectly p -decodable μ -secure erasure coding scheme. Partial decoding starts when the first $\mu + p$ nodes are accessed. At this point, the first subset of p file symbols (s_1, \dots, s_p) is reconstructed. When the first $\mu + 2p$ nodes are accessed, the next subset of p file symbols (s_{p+1}, \dots, s_{2p}) can also be decoded. In fact, if the user requests only (s_{p+1}, \dots, s_{2p}), it is sufficient to access the first μ nodes and the nodes numbered from $\mu + p + 1$ to $\mu + 2p$. The general requirement for partial decoding is: ($s_{rp+1}, \dots, s_{(r+1)p}$) can be reconstructed by accessing the first μ nodes and p additional nodes numbered from $\mu + rp + 1$ to $\mu + (r + 1)p$. We always assume that $p|(k - \mu)$.

In this work we investigate a construction of erasure coding schemes for DSS that are secure in terms of data confidentiality: even when some of its storage nodes are compromised or eavesdropped by some unwanted party, the stored file is still kept confidential. Moreover, such secure codes must provide easy partial decoding for a legitimate user, which can often access more nodes than an illegal adversary. We proposed the so-called p -decodable μ -secure erasure coding scheme ($1 \leq p \leq k - \mu, 0 \leq \mu < k$), which satisfies the MDS property (the file can be reconstructed by accessing any k out of n storage nodes) and the following additional properties:

- (P1) **strongly secure up to a threshold**: an adversary which eavesdrops at most μ storage nodes gain no information (in Shannon's sense) about the stored file,
- (P2) **partially decodable**: a legitimate user can recover a subset of p file symbols by accessing some $\mu + p$ storage nodes.

Regarding (P2), throughout this paper we always assume that $p|(k - \mu)$. In other words, we can always partition the set of $k - \mu$ file symbols into subsets of size p each. Apart from (P1)-(P2), if the following additional property is also satisfied, the scheme is referred to as *perfectly* p -decodable μ -secure:

- (P3) **weakly secure up to a threshold**: an adversary which eavesdrops more than μ but less than $\mu + p$ storage nodes cannot reconstruct any part of the file,

We illustrate the properties of a perfect coding scheme in Fig 1.

	Code rate	Strong security threshold	Weak security threshold	Partial decodability threshold
Systematic erasure code	k/n	0	0	1
Codes for (erasure) wiretap channel II [6], [7]	$(k - \mu)/n$	μ	N.A.	N.A.
Ramp secret sharing scheme [8], [9], [10]	$(k - \mu)/n$	μ	$k - 1$	k
p-decodable μ-secure code	$(k - \mu)/n$	μ	N.A.	$\mu + p$
Perfectly p-decodable μ-secure code	$(k - \mu)/n$	μ	$\mu + p - 1$	$\mu + p$

TABLE I: Comparison among five erasure coding schemes. Strong (weak) security threshold refers to the maximum number of storage nodes that the adversary is allowed to access without jeopardizing the strong (weak) security of the scheme. Partial decodability threshold refers to the number of nodes an user has to access to start decoding the file partially. Here $0 \leq \mu \leq k - 1$, $1 \leq p \leq k - \mu$, and $p | (k - \mu)$. When $p = 1$ and $\mu = 0$, the (perfectly) p -decodable μ -secure scheme is actually systematic. An ‘N.A.’ entry means that the corresponding threshold can take any value and that threshold is not even considered in the design of the coding scheme.

Note that when $p = k - \mu$, only (strong and weak) security is guaranteed and there is no partial decodability: an user cannot retrieve any part of the file by accessing less than k nodes. Such a secure coding scheme was first studied in the work of Yamamoto [8] in the context of ramp secret sharing scheme. Recently, superregular matrices (all square submatrices are invertible) such as Cauchy matrices have also been employed to construct such codes [9], [10]. Similar work in secure regenerating codes, which can be regarded as vector erasure codes with optimal node repair, can be found in [11], [12].

In this work we address the gap in the literature regarding partial decodability of erasure coding schemes. In our proposed coding scheme, while Property (P1) and Property (P3) provide (strong and weak) security, Property (P2) guarantees partial decodability: once the user accesses p more nodes than the strong security threshold μ , it can start to decode some subset of p file symbols. A secure erasure code that supports partial decoding is particularly useful in applications involving retrieval of large files. A typical example is in video streaming services, where a large-size video is often split into chunks and these chunks are then streamed one by one to the user. Our proposed coding scheme, if employed in such services, would provide not only confidentiality but also ease of partial retrieving of the video to any desired level. We stress that extending the existing secure coding schemes by taking into account partial decodability does not result in any overhead.

	Small field	Large field
p -decodable μ -secure	$*\forall \mu, \forall p$	N.A.
perfectly p -decodable μ -secure	$p = k - \mu$ [8], [9], [10] $*p = 1, \forall \mu$ $*\mu = 0, 1, \forall p$	$*\forall \mu, \forall p$

TABLE II: A summary of existence of (perfectly) p -decodable μ -secure coding schemes. The entries with an asterisk “*” are the new results established in this work. A field is “small” if its size is a polynomial in n and k , or “large” if otherwise.

Our main contribution is summarized below.

- We provide an explicit construction over small fields for p -decodable μ -secure coding schemes for any p and μ .
- We provide an explicit construction over small fields for *perfectly* p -decodable μ -secure coding schemes when $p = 1$ (for every μ), and when $\mu = 0, 1$ (for every $p, p | k - \mu$).

- We prove that perfectly p -decodable μ -secure schemes exist over sufficiently large fields for almost all p and μ .

The paper is organized as follows. Section II provides necessary notation and definitions. The construction of p -decodable μ -secure schemes is presented in Section III. We discuss the existence of perfect coding schemes in Section IV. The paper is concluded in Section V.

II. PRELIMINARIES

A. Notation

Let \mathbb{F}_q denote the finite field with q elements. Let $[n]$ denote the set $\{1, 2, \dots, n\}$. For a vector $\mathbf{u} = (u_1, \dots, u_n) \in \mathbb{F}_q^n$, let \mathbf{u}^T denote the transpose of \mathbf{u} . For any $n \geq 1$ let \mathbf{I}_n denote the identity matrix of order n . We also use \mathbf{O} to denote the all-zero matrix of certain size.

Below we define standard notation from coding theory (see [1]). The (Hamming) *weight* of \mathbf{u} is $\text{wt}(\mathbf{u}) = |\{i : u_i \neq 0\}|$. The (Hamming) *distance* between two vectors \mathbf{u} and \mathbf{v} is $d(\mathbf{u}, \mathbf{v}) = \text{wt}(\mathbf{u} - \mathbf{v})$. An $[n, k, d]_q$ (error-correcting) code \mathcal{C} (sometimes d and q are dropped) is a subspace of the vector space \mathbb{F}_q^n of dimension k such that the minimum distance between any two distinct vectors in that subspace is at least d . We called $d = d(\mathcal{C})$ the *minimum distance* of the code. The well-known Singleton bound states that for any $[n, k, d]_q$ code, $d \leq n - k + 1$. A code attaining the Singleton bound is called maximum distance separable (MDS). Any vector in a code is referred to as a codeword. A generator matrix of an $[n, k]_q$ code is a $k \times n$ matrix whose rows are linearly independent codewords of the code.

B. Coset Coding Technique

Let $\mathbf{r} = (r_1, \dots, r_\mu)$ be a vector of independent and identically uniformly distributed random variables over \mathbb{F}_q . Let $\mathbf{S} = (S_1, S_2, \dots, S_{k-\mu})$, where S_i ’s ($i \in [k - \mu]$) are independent and identically uniformly distributed random variables over some alphabet \mathbb{F}_q . We assume that the file to be stored in the system is $\mathbf{s} = (s_1, s_2, \dots, s_{k-\mu}) \in \mathbb{F}_q^{k-\mu}$, a realization of \mathbf{S} . We call $k - \mu$ the file size and each s_i a file symbol.

We denote by $\mathcal{D}(n, k)$ a DSS with n storage nodes where the file can be recovered from the contents of any k out of n nodes. Node i ($i \in [n]$) stores a coded symbol c_i , which is a function of the file symbols s_i ’s and the random symbols r_j ’s. Let $\mathbf{c} = (c_1, c_2, \dots, c_n)$. Let $\mathbf{C} = (C_1, C_2, \dots, C_n)$ be

\mathbf{c} 's corresponding vector of random variables over \mathbb{F}_q . We only consider here scalar linear erasure coding schemes, based on $[n, k]_q$ MDS codes, described as follows. The file $\mathbf{s} \in \mathbb{F}_q^{k-\mu}$ is encoded to $\mathbf{c} = \mathbf{x}\mathbf{G} \in \mathbb{F}_q^n$, where

- $\mathbf{x} = (\mathbf{s} \mid \mathbf{r})$ is obtained by concatenating \mathbf{s} and \mathbf{r} ,
- \mathbf{G} is a generator matrix of an $[n, k, n - k + 1]_q$ MDS code. We often refer to \mathbf{G} as the *coding matrix*.

It is well known in coding theory [1] that if Node i stores c_i produced as above then the file can be recovered by accessing any k nodes. For the coding scheme to be strongly secure against an adversary that can access μ nodes (see Definition 1), the last μ rows of \mathbf{G} must generate an $[n, \mu]$ MDS code. In fact, this is an equivalent way to describe the coset coding technique invented by Ozarow and Wyner [6]. This technique has been widely adopted in the network coding literature to secure a network code against a wiretapper (see, for instance [13] and references therein).

C. Security and Partial Decodability

We assume that an adversary can eavesdrop/access any m storage nodes of its choice and tries to learn illegally the content of the stored file. We refer to m as the adversary's *strength*. In the following definitions, recall that $\mathbf{S} = (S_1, S_2, \dots, S_{k-\mu})$ represents the file stored in the system.

Definition 1. An erasure coding scheme for a DSS $\mathcal{D}(n, k)$ is called *strongly secure* against an adversary of strength m ($m < k$) if the entropy

$$H(\mathbf{S} \mid \{C_i : i \in E\}) = H(\mathbf{S}),$$

for all subsets $E \subseteq [n]$, $|E| \leq m$. We also refer to such a coding scheme as *strongly m -secure*.

In words, a coding scheme is strongly μ -secure if an adversary which can access an arbitrary set of at most μ storage nodes cannot obtain any information at all about the stored file. It is well-known that as long as the bottom $\mu \times n$ submatrix of \mathbf{G} also generates an $[n, \mu]$ MDS code then the coding scheme described in Section II-B is strongly μ -secure. The MDS code generated by such a matrix \mathbf{G} is often called a *nested* MDS code in the literature.

Definition 2. An erasure coding scheme for a DSS $\mathcal{D}(n, k)$ is called *weakly secure* against an adversary of strength m ($m < k$) if the entropy

$$H(S_j \mid \{C_i : i \in E\}) = H(S_j),$$

for all $j \in [k - \mu]$ and for all subsets $E \subseteq [n]$, $|E| \leq m$. We also refer to such a coding scheme as *weakly m -secure*.

The following lemma specifies a necessary and sufficient condition for the weak security of the erasure coding scheme described in Section II-B.

Lemma 3. Let $\mathbf{s} = (s_1, s_2, \dots, s_{k-\mu}) \in \mathbb{F}_q^{k-\mu}$ be the stored file and $\mathbf{r} = (r_1, r_2, \dots, r_\mu)$ be some random vector over \mathbb{F}_q . Let $(\mathbf{s} \mid \mathbf{r})\mathbf{M}$, where \mathbf{M} is a $k \times m$ matrix over \mathbb{F}_q , represent

the m coded symbols stored at some m storage nodes that the adversary has access to. Then the adversary cannot determine any particular file symbol s_i if and only if the column space of \mathbf{M} does not contain \mathbf{e}_i for every $i \in [k - \mu]$, where \mathbf{e}_i is the unit vector with only one nonzero coordinate at the i th position.

Proof: Appendix A. ■

The intuition behind the proof of this lemma is explained below. As the adversary obtains $(\mathbf{s} \mid \mathbf{r})\mathbf{M}$, it can linearly transform these coded symbols by considering the product $(\mathbf{s} \mid \mathbf{r})\mathbf{M}\boldsymbol{\alpha}^T$, where $\boldsymbol{\alpha} \in \mathbb{F}_q^m$ is some coefficient vector. The adversary can determine a file symbol s_i ($i \in [k - \mu]$) if and only if there exists $\boldsymbol{\alpha}$ so that $\mathbf{M}\boldsymbol{\alpha}^T = \mathbf{e}_i$. As $\mathbf{M}\boldsymbol{\alpha}^T$ is a vector of \mathbb{F}_q^m , we derive the conclusion in Lemma 3.

After Yamamoto [8], the concept of weak security was also discovered by Bhattad and Narayanan [14] in a more general context of network coding. Weak security is important in practice since it guarantees that no meaningful information is leaked to the adversary, and often requires no additional overhead. For instance, suppose that the file $\mathbf{s} = (s_1, s_2, s_3)$ is to be stored in a DSS with four storage nodes, which tolerates one node failure. Using an usual systematic erasure code, the four nodes store $\mathbf{c} = (s_1, s_2, s_3, s_1 + s_2 + s_3)$. However, if an adversary can access any node among the first three, then it can retrieve some s_i , which is part of the file. On the other hand, if the file is encoded into $\mathbf{c} = (s_1 + s_2, s_1 + s_3, s_2 + s_3, s_1 + s_2 + s_3)$, then an adversary who accesses one storage node would not be able to determine any s_i . Indeed, for instance if it observes $s_1 + s_2$, then it cannot determine the exact value of either s_1 or s_2 , as for the adversary, both s_1 and s_2 are completely random variables. If \mathbf{s} is a video and s_i 's are movie chunks, then by using the latter coding scheme, an adversary who observes one storage node cannot determine each chunk, and hence, cannot play any part of the movie. Such coding scheme is said to be weakly secure against an adversary of strength one, or weakly 1-secure. Moreover, that coding scheme consists of the same number of storage nodes and can also tolerate one node failure, hence introduces no storage overhead compared to a normal systematic code. In fact, while strong security always comes with a cost in storage capacity, weak security is often given for free.

Definition 4. We consider the coding scheme described in Section II-B, where \mathbf{G} is a generator matrix of an $[n, k, n - k + 1]_q$ MDS code and the file $\mathbf{s} \in \mathbb{F}_q^{k-\mu}$ is encoded into the coded vector $\mathbf{c} = (\mathbf{s} \mid \mathbf{r})\mathbf{G} \in \mathbb{F}_q^n$. Suppose that $0 \leq \mu < k$, $1 \leq p \leq k - \mu$, and moreover $p \mid (k - \mu)$. The coding scheme based on \mathbf{G} is *p -decodable μ -secure* if it satisfies the following properties simultaneously.

- (P1) It is strongly μ -secure as defined in Definition 1.
- (P2) It is p -decodable: each subset of p file symbols $\{s_{rp+1}, s_{rp+2}, \dots, s_{(r+1)p}\}$ ($0 \leq r \leq (k - \mu)/p - 1$) can be reconstructed from the content of some $\mu + p$ storage nodes.

The coding scheme is *perfectly p -decodable μ -secure* if it satisfies the following additional property:

(P3) It is weakly $(\mu + p - 1)$ -secure as defined in Definition 2. We also say that the corresponding coding matrix \mathbf{G} is (perfectly) p -decodable μ -secure.

Remark 5.

- A (perfectly) 1-decodable 0-secure code is simply a systematic code in the classical sense, where each file symbol is stored as is (in its clear form) at some node.
- A perfectly $(k - \mu)$ -decodable μ -secure code is the type of secure codes studied in the work of Yamamoto [9] and Olivera *et al.* [10].
- We can replace the weak security in (P3) by the *block security* (equivalently, security against guessing) [14], [11] which requires a stronger condition that no information about any *subset* of file symbols up to a certain size is known to the adversary. In fact, all of our results in this work can be extended to block security. However, to simplify the presentation, we restrict ourselves to only weak security.

We illustrate the concept of perfectly p -decodable μ -secure coding scheme in the following example.

Example 6. Let $k = 5$, $n = 6$, $\mu = 1$, $p = 2$, and $q = 11$. Let $\mathbf{s} = (s_1, s_2, s_3, s_4) \in \mathbb{F}_{11}^4$ be the stored file and r a random symbol over \mathbb{F}_{11} . We use the following coding matrix

$$\mathbf{G} = \begin{pmatrix} 0 & 1 & 6 & 0 & 0 & 7 \\ 0 & 6 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 10 & 7 \\ 0 & 0 & 0 & 7 & 1 & 1 \\ 4 & 5 & 10 & 1 & 9 & 6 \end{pmatrix}$$

The coding scheme is

$$\mathbf{c} = (\mathbf{s} \mid r)\mathbf{G}.$$

We show below that this coding scheme is perfectly 2-decodable 1-secure. Firstly, it is easy to verify that \mathbf{G} generates a $[6, 5]_{11}$ MDS code. Therefore, the file can be reconstructed by accessing any five storage nodes. The bottom 1×6 matrix also obviously generates a $[6, 1]$ MDS code, hence guarantees that the scheme is strongly 1-secure. Hence (P1) is satisfied. To verify (P3), note that $\mu + p - 1 = 1 + 2 - 1 = 2$. We can easily verify that any two columns of \mathbf{G} do not generate an unit vector \mathbf{e}_i for every $i \in [4]$. Hence, according to Lemma 3, the coding scheme is weakly 2-secure. Lastly, for (P2), we prove that each of the 2-subsets $\{s_1, s_2\}$ and $\{s_3, s_4\}$ can be reconstructed by accessing some three nodes ($\mu + p = 3$). Indeed, by accessing the first three nodes, an user obtains the product

$$(s_1, s_2, s_3, s_4, r) \begin{pmatrix} 0 & 1 & 6 \\ 0 & 6 & 4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 4 & 5 & 10 \end{pmatrix}$$

If we post-multiply the above product with the vector $\alpha = (10, 4, 5)^T$ then we obtain s_1 . If we post-multiply the above product with the vector $\beta = (5, 5, 1)^T$ then we obtain s_2 . Hence, $\{s_1, s_2\}$ can be reconstructed by accessing the first three nodes. Similarly, we can verify that $\{s_3, s_4\}$ can be reconstructed by accessing Node 1, Node 4, and Node 5.

III. A CONSTRUCTION OF p -DECODABLE μ -SECURE CODING SCHEMES

We establish in this section a general construction for p -decodable μ -secure erasure coding schemes, which satisfy (P1) and (P2): an adversary which can access the content of μ storage nodes gains no information about stored file, and a legitimate user can retrieve each subset of p file symbols by accessing some $\mu + p$ storage nodes.

A. A General Construction

We start with the definition of superregular matrices, which is critical in our construction.

Definition 7 ([15]). A *superregular* matrix is a matrix where every square submatrix is invertible.

Two well-known constructions of superregular matrices are via Cauchy matrices and Vandermonde matrices [16]. A Cauchy matrix is a matrix of the form $\mathbf{C} = (1/(x_i - y_j))_{i,j}$ where x_i 's and y_j 's are distinct elements of any finite field \mathbb{F}_q . Cauchy matrices are superregular by themselves. The following straightforward results about superregular matrices are especially useful in our construction.

Lemma 8. Let \mathbf{C} be a superregular $k \times n$ matrix ($k \leq n$). Then the following hold.

- Any submatrix of \mathbf{C} is also superregular.
- Any subset of k' rows of \mathbf{C} ($1 \leq k' \leq k$) generates an $[n, k']$ MDS code. Hence, every nontrivial vector generated by these k' rows has weight at least $n - k' + 1$.
- Any subset of n' columns of \mathbf{C} ($1 \leq n' \leq k$), generates a $[k, n']$ MDS code. Hence, every nontrivial vector generated by these n' columns has weight at least $k - n' + 1$.

We now describe our general construction, using the so-called *partial superregular* matrices.

Main Construction.

- **Step 1.** Choose any superregular $k \times n$ matrix \mathbf{G}'' and write it in the following block form.

$$\mathbf{G}'' = \left(\begin{array}{c|c|c} \mathbf{A}'' & \mathbf{B}'' & \mathbf{C}'' \\ \hline \mathbf{D} & \mathbf{E} & \mathbf{F} \end{array} \right)_{\mu}^{k-\mu, n-k} \quad (1)$$

- **Step 2.** Perform elementary row operations to turn the matrix \mathbf{A}'' at the top-left corner into an all-zero matrix. We can do so by adding certain linear combination of the last μ rows of \mathbf{G}'' to each of its first $k - \mu$ rows. Note that the $\mu \times \mu$ matrix \mathbf{D} is invertible, hence its rows can generate any vector of length μ . The resulting matrix, referred to as \mathbf{G}' , can be presented in block form as below. Since there is no row operation performed on

the last μ rows, the three block submatrices D , E , F are the same in G'' and G' .

$$G' = \left(\begin{array}{c|c|c} \mu & k-\mu & n-k \\ \hline \mathbf{O} & \mathbf{B}' & \mathbf{C}' \\ \hline \mathbf{D} & \mathbf{E} & \mathbf{F} \end{array} \right)_{\mu}^{k-\mu} \quad (2)$$

- **Step 3.** Perform elementary row operations on the first $k - \mu$ rows of G' to turn the square submatrix B' into a new square matrix B of the same size determined as follows. It is a block diagonal matrix where each block submatrix is of size $p \times p$. Moreover, except from the zero entries, all other entries, which belong to those block $p \times p$ submatrices, are the same as the corresponding entries in B'' . Equivalently, B can be obtained from B'' by turning those entries that do not belong to any block diagonal submatrix into zero. We can write B in the block form as below, where B''_i ($1 \leq i \leq (k - \mu)/p$) is the i th diagonal block $p \times p$ submatrix of B'' .

$$B = \begin{pmatrix} p & p & \cdots & p \\ \mathbf{B}''_1 & & & \\ & \mathbf{B}''_2 & & \\ & & \ddots & \\ & & & \mathbf{B}''_{\frac{k-\mu}{p}} \end{pmatrix} \begin{matrix} p \\ p \\ \vdots \\ p \end{matrix} \quad (3)$$

Such transformation can always be done because both B'' and B are invertible. Thus, the coding matrix G , as the output of Step 3, is determined by

$$G = TG',$$

where the transform matrix T is

$$T = \left(\begin{array}{c|c} k-\mu & \mu \\ \hline \mathbf{B}\mathbf{B}'^{-1} & \mathbf{O} \\ \hline \mathbf{O} & \mathbf{I}_{\mu} \end{array} \right)_{\mu}^{k-\mu}$$

We can write G in block format as follows.

$$G = \begin{pmatrix} \mu & p & p & \cdots & p & n-k \\ \hline \begin{matrix} p \\ p \\ \vdots \\ p \end{matrix} & \begin{matrix} \mathbf{B}''_1 \\ \mathbf{B}''_2 \\ \mathbf{O} \\ \mathbf{B}''_{\frac{k-\mu}{p}} \end{matrix} & \mathbf{O} & & & \mathbf{C} \\ \hline \mathbf{D} & \mathbf{E} & \mathbf{F} & & & \end{pmatrix} \begin{matrix} k-\mu \\ \mu \end{matrix}$$

where $C = BB'^{-1}C'$.

Note that the coding matrix G produced by the Main Construction has the same entries as the original superregular matrix G'' , except for those zero entries and entries in block submatrix C . Therefore, we often refer to G constructed in this way as *partial* superregular matrix. We illustrate the steps to construct a partial superregular matrix in Example 9.

Example 9. Let $k = 5$, $n = 6$, $q = 11$, $\mu = 1$, and $p = 2$.

Step 1. Let us choose G'' to be a Cauchy matrix

$$G'' = \left(\begin{array}{c|ccc|c} 2 & 1 & 6 & 3 & 7 & 9 \\ 8 & 6 & 4 & 9 & 5 & 2 \\ 7 & 4 & 3 & 2 & 10 & 8 \\ 10 & 9 & 2 & 7 & 1 & 5 \\ \hline 4 & 5 & 10 & 1 & 9 & 6 \end{array} \right).$$

Step 2. As the bottom-left entry is nonzero, we can add certain multiple of the last row to each of the first four rows of G'' to obtain

$$G' = \left(\begin{array}{c|ccc|c} 0 & 4 & 1 & 8 & 8 & 6 \\ 0 & 7 & 6 & 7 & 9 & 1 \\ 0 & 9 & 2 & 3 & 8 & 3 \\ 0 & 2 & 10 & 10 & 6 & 1 \\ \hline 4 & 5 & 10 & 1 & 9 & 6 \end{array} \right).$$

Step 3. Finally, we can perform certain elementary row operations on the first four rows to obtain the coding matrix

$$G = \left(\begin{array}{c|ccc|c} 0 & 1 & 6 & 0 & 0 & 7 \\ 0 & 6 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 10 & 7 \\ 0 & 0 & 0 & 7 & 1 & 1 \\ \hline 4 & 5 & 10 & 1 & 9 & 6 \end{array} \right),$$

which is proved earlier in Example 6 to generate a (perfectly) 2-decodable 1-secure coding scheme.

The following lemmas assert that a coding scheme based on a partial superregular matrix, which is produced by the Main Construction, is indeed p -decodable μ -secure.

Lemma 10. *The partial superregular matrix G produced by the Main Construction generates an $[n, k]$ MDS code. Moreover, the last μ rows of G also generates an $[n, \mu]$ MDS code. As a consequence, the coding scheme based on G has the MDS property and moreover, satisfies (P1) (i.e., being strongly μ -secure).*

Proof: As G'' generates an $[n, k]$ MDS code and G is obtained from G'' by applying only elementary row operation, G generates the same $[n, k]$ MDS code. Since the last μ rows of G are the same as those of G'' , they also generates an $[n, \mu]$ MDS code. Hence (P1) is satisfied. ■

Lemma 11. *An erasure coding scheme based on a partial superregular matrix G (Main Construction) always satisfies (P2): each subset of p file symbols $\{s_{rp+1}, s_{rp+2}, \dots, s_{(r+1)p}\}$ ($0 \leq r \leq (k - \mu)/p - 1$) can be reconstructed from the content of some $\mu + p$ storage nodes.*

Proof: Recall that the file s is encoded into $c = (s \mid r)G$. We prove that the first p file symbols $\{s_1, s_2, \dots, s_p\}$ can be reconstructed from the first $\mu + p$ coordinates of c . In general, it can be proved in a similar manner that for each r where $0 \leq r \leq (k - \mu)/p - 1$, the p file symbols $\{s_{rp+1}, s_{rp+2}, \dots, s_{(r+1)p}\}$ can be reconstructed from $\mu + p$ coordinates of c , namely c_1, \dots, c_{μ} , and $c_{\mu+rp+1}, \dots, c_{\mu+(r+1)p}$.

The coding matrix G produced by the Main Construction can be presented in the block form as follows.

$$G = \left(\begin{array}{c|cc|c|c|c|c} & B_1'' & & & & \\ & & B_2'' & & & \\ & & & \ddots & & \\ & & & & B_{(k-1)/p}'' & \\ D & E_1 & E_2 & \cdots & E_{(k-1)/p} & F \end{array} \right),$$

where each E_i is a $\mu \times p$ submatrix of E such that $E = (E_1 | E_2 | \cdots | E_{(k-1)/p})$. According to the structure of G , the first $\mu + p$ coordinates of c can be written as

$$(c_1 \cdots c_{\mu+p}) = (s_1 \cdots s_p | r_1 \cdots r_\mu) H_1,$$

where

$$H_1 = \left(\begin{array}{c|c} \overset{\mu}{O} & \overset{p}{B_1''} \\ D & E_1 \end{array} \right)_\mu.$$

Since G'' is superregular, both B_1'' and D are invertible. Let $v \in \mathbb{F}_q^p$ be the i th column in the inverse of B_1'' ($1 \leq i \leq p$). Moreover, let $u \in \mathbb{F}_q^\mu$ be the column vector satisfying $Du = -E_1 v$. Then

$$H_1 \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} B_1'' v \\ Du + E_1 v \end{pmatrix} = \begin{pmatrix} e_i \\ 0 \end{pmatrix},$$

where

$$e_i = (\underbrace{0, \dots, 0}_{i-1}, 1, \underbrace{0, \dots, 0}_{p-i})^T.$$

Therefore, the user can reconstruct s_i as follows.

$$\begin{aligned} s_i &= (s_1 \cdots s_p | r_1 \cdots r_\mu) \begin{pmatrix} e_i \\ 0 \end{pmatrix} \\ &= (s_1 \cdots s_p | r_1 \cdots r_\mu) H_1 \begin{pmatrix} u \\ v \end{pmatrix} \\ &= (c_1 \cdots c_{\mu+p}) \begin{pmatrix} u \\ v \end{pmatrix}. \end{aligned}$$

As i can be chosen arbitrarily between 1 and p , we conclude that an user which has access to the first $\mu + p$ coordinates of c can reconstruct all s_i for $1 \leq i \leq p$. The proof follows. ■

Theorem 12. *The Main Construction produces a coding matrix that generates a p -decodable μ -secure coding scheme for all $0 \leq \mu < k$ and $1 \leq p \leq k - \mu$, $p|(k - \mu)$.*

Proof: According to Lemma 10 and Lemma 11, the Main Construction produces a coding matrix that generates a coding scheme satisfying both (P1) and (P2). Hence, such scheme is p -decodable μ -secure, according to Definition 4. ■

IV. ON PERFECTLY p -DECODABLE μ -SECURE CODING SCHEMES

In this section, we first prove that a coding matrix produced by the Main Construction in Section III-A also satisfies (P3) when $p = 1$ ($\forall \mu$) and when $\mu = 0, 1$ ($\forall p$). Finally, we establish the existence of perfectly p -decodable μ -secure coding schemes over sufficiently large fields for almost every p and μ (namely, $k \geq 2(\mu + p) - 1$).

A. The Case $p = 1$

A (perfectly) 1-decodable μ -secure coding scheme is the best scheme among all strongly μ -secure schemes in terms of partial decoding. Such a scheme allows the user to reconstruct one file symbol right after the user accesses one more node than the security threshold μ . It is a sharp turn from knowing nothing about the file to being able to reconstruct one file symbol. In fact, according to Lemma 11, after accessing the first μ nodes, accessing any additional node would give the user one new file symbol. Hence, beyond the threshold μ , the coding scheme works in a similar manner as the conventional systematic coding scheme. Note that when $p = 1$ and $\mu = 0$, a (perfectly) 1-decodable 0-secure coding scheme is nothing other than a normal systematic coding scheme.

Lemma 13. *When $p = 1$, the Main Construction produces a matrix G that generates a (perfectly) 1-decodable μ -secure coding scheme for any $\mu \geq 0$.*

Proof: When $p = 1$, (P3) is satisfied trivially. Hence, the Main Construction yields a coding scheme satisfying simultaneously (P1), (P2), and (P3). Such a coding scheme, according to Definition 4, is perfectly 1-decodable μ -secure. ■

B. The Case $\mu = 0$

A perfectly p -decodable 0-secure coding scheme is of particular interest because of the following properties.

- **Weak security with no overhead on storage capacity:** the scheme provides weak security against an adversary which can access up to $p - 1$ storage nodes. Such an adversary cannot reconstruct any part of the stored file. Moreover, no storage overhead occurs compared to a normal erasure coding scheme: the file size is k and the code is an $[n, k]$ MDS code. That is because there are no random symbols employed in the scheme.
- **p -partial decodability:** the user can reconstruct each subset of p file symbols by accessing certain p storage nodes.

Lemma 14. *When $\mu = 0$, the Main Construction produces a matrix G that generates a perfectly p -decodable 0-secure coding scheme for any $p \geq 1$, $p|k$.*

Proof: Due to Lemma 10 and Lemma 11, it suffices to prove that the coding matrix G produced by the Main Construction generates a coding scheme satisfying (P3) - weak security. According to Lemma 3, we aim to show that any set of $p - 1$ ($= \mu + p - 1$) columns of G does not generate an unit vector of weight one. As $p|k$, we consider the following two cases: $k = p$ and $k \geq 2p$.

If $k = p$, as $\mu = 0$, the coding matrix G is simply the same as the input superregular matrix G'' . By Lemma 8, any set of $p - 1$ columns of G generates a $[p, p - 1, 2]$ MDS code, hence never generates an unit vector e_i , which has weight less than two. Therefore, according to Lemma 3, the coding scheme based on G is weakly secure against an adversary which can access at most $p - 1$ nodes. Hence (P3) is satisfied.

We now assume that $k \geq 2p$. When $\mu = 0$, the matrix \mathbf{G} has the following form.

$$\mathbf{G} = \left(\begin{array}{ccc|c} \mathbf{B}_1'' & & & \mathbf{C} \\ & \mathbf{B}_2'' & & \\ & & \ddots & \\ & & & \mathbf{B}_{k/p}'' \end{array} \right).$$

We assume, by contradiction, that some set L of $p-1$ columns of \mathbf{G} generates an unit vector $\mathbf{e}_i \in \mathbb{F}_q^k$. Without loss of generality, we can assume that $i = 1$. For simplicity, we slightly abuse the notation and also use L to denote the set of indices of the columns in L . Then there exist some coefficients $\alpha_j \in \mathbb{F}_q$ ($j \in L$), so that

$$\sum_{j \in L} \alpha_j \mathbf{G}[j] = \mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (4)$$

where $\mathbf{G}[j]$ denotes the j th column of \mathbf{G} . Note that the $p \times p$ block matrix \mathbf{B}_1'' is a square submatrix of a superregular matrix \mathbf{G}'' , and hence is invertible. Therefore, there exists a linear combination of the first p columns of \mathbf{G} that generate the vector $-\mathbf{e}_1$. In other words, there exist some coefficients $\beta_j \in \mathbb{F}_q$ ($j \in [p]$), such that

$$\sum_{j=1}^p \beta_j \mathbf{G}[j] = -\mathbf{e}_1 = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}. \quad (5)$$

By (4) and (5), we deduce that

$$\sum_{j=1}^p \beta_j \mathbf{G}[j] + \sum_{j \in L} \alpha_j \mathbf{G}[j] = \mathbf{0}. \quad (6)$$

Hence, there exists a linear combination of these at most $(2p-1)$ columns of \mathbf{G} that is equal to $\mathbf{0}$. Note that we assume $k \geq 2p$. Moreover, according to Lemma 10, \mathbf{G} is an $[n, k]$ MDS code. Hence, any set of at most k columns of \mathbf{G} must be linear independent. We obtain below a contradiction by arguing that the linear combination of at most $2p-1 < k$ columns in (6) is *nontrivial*, that is, their coefficients are not identically zero. Then the proof would follow.

First, according to Lemma 8, as \mathbf{B}_1'' is superregular, any $p-1$ columns of \mathbf{B}_1'' form a $[p, p-1, 2]$ MDS code. In other words, any $p-1$ columns of \mathbf{B}_1'' does not generate a nonzero vector of weight less than two. Combining this fact with (5), we deduce that $\beta_j \neq 0$ for all $j \in [p]$. Therefore, in the linear combination (6), while the first sum consists of p columns of \mathbf{G} with all nonzero coefficients, the second sum is a linear combination of $p-1$ columns of \mathbf{G} . Hence, there must be at least one term in the first sum that cannot be canceled out. Thus, (6) is a *nontrivial* linear combination of less than k columns of \mathbf{G} . This conclusion contradicts the fact that any k or less columns of \mathbf{G} must be linearly independent. ■

C. The Case $\mu = 1$

Lemma 15. When $\mu = 1$, the Main Construction produces a matrix \mathbf{G} that generates a perfectly p -decodable 1-secure coding scheme for any $p \geq 1$, $p|(k-1)$.

Proof: Due to Lemma 10 and Lemma 11, it suffices to prove that the coding matrix \mathbf{G} produced by the Main Construction generates a coding scheme satisfying (P3) - weak security. According to Lemma 3, we aim to show that any set of p ($= \mu + p - 1$) columns of \mathbf{G} does not generate an unit vector of weight one. As $p|(k-1)$, there are two cases: $k-1 = p$ and $k-1 \geq 2p$.

In the case $k-1 = p$, there is no partial decodability and we simply use the superregular matrix \mathbf{G}'' in Step 1 of the Main Construction, instead of \mathbf{G} . By Lemma 8, any set of $p = k-1$ columns of \mathbf{G}'' generates a $[p, p-1, 2]$ MDS code, hence cannot generate an unit vector \mathbf{e}_i of weight one. Therefore, the coding scheme based on \mathbf{G}'' is weakly secure against an adversary which can access at most $p-1$ storage nodes. Hence (P3) is satisfied.

Now we assume that $k-1 \geq 2p$. When $\mu = 1$, the matrix \mathbf{G} has the following form.

$$\mathbf{G} = \left(\begin{array}{ccc|c} \mathbf{B}_1'' & & & \mathbf{C} \\ \mathbf{0} & \mathbf{B}_2'' & & \\ & & \ddots & \\ & & & \mathbf{B}_{(k-1)/p}'' \end{array} \middle| \begin{array}{c} \mathbf{C} \\ \mathbf{D} \\ \mathbf{E}_1 \\ \mathbf{E}_2 \\ \vdots \\ \mathbf{E}_{(k-1)/p} \end{array} \right),$$

where $\mathbf{0}$ is a $(k-1) \times 1$ all-zero column vector, $\mathbf{D} = (d_{1,1})$ is a 1×1 matrix, and \mathbf{E}_i is a $1 \times p$ row vector for each $1 \leq i \leq (k-1)/p$. Note that $\mathbf{E} = (\mathbf{E}_1 | \mathbf{E}_2 | \cdots | \mathbf{E}_{(k-1)/p})$.

We now assume, by contradiction, that some set L of p columns of \mathbf{G} generates an unit vector $\mathbf{e}_i \in \mathbb{F}_q^k$, for some $i \in [k-1]$. Without loss of generality, we can assume that $i = 1$. For simplicity, we slightly abuse the notation and also use L to denote the set of indices of the columns in L . Then there exist some coefficients $\alpha_j \in \mathbb{F}_q$ ($j \in L$), so that

$$\sum_{j \in L} \alpha_j \mathbf{G}[j] = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (7)$$

where $\mathbf{G}[j]$ denotes the j th column of \mathbf{G} .

Let $J = \{2, 3, \dots, p+1\}$. We first argue that $L \neq J$. Indeed, the matrix

$$\left(\begin{array}{c} \mathbf{B}_1'' \\ \mathbf{E}_1 \end{array} \right) \quad (8)$$

is a submatrix of the superregular matrix \mathbf{G}'' chosen in Step 1 of the Main Construction. Therefore, according to Lemma 8, its columns generate a $[p+1, p, 2]$ MDS code. Hence, every nontrivial linear combination of these p columns has weight at

least two. As a consequence, any nontrivial linear combination of p columns $\mathbf{G}[j]$'s ($j \in J$), which correspond to the columns of the matrix in (8), has weight at least two. However, due to (7), the columns in L can generate the unit vector \mathbf{e}_1 , which has weight one. Thus, L and J must be different.

As \mathbf{G}'' is superregular, \mathbf{B}_1'' is invertible. Therefore, there exist some coefficients $\beta_j \in \mathbb{F}_q$ ($j \in J$), such that

$$\sum_{j \in J} \beta_j \mathbf{G}[j] = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ \hline \mathbf{E}_1 \mathbf{\beta}^T \end{pmatrix}, \quad (9)$$

where $\mathbf{\beta} = (\beta_2, \dots, \beta_{p+1})$. According to Lemma 8, any set of less than p columns of \mathbf{B}_1'' would never generate a nontrivial vector of weight less than two. Therefore, from (9), we deduce that $\beta_j \neq 0$ for every $j \in J$. Moreover, since \mathbf{G}'' is superregular, $d_{1,1}$ must be nonzero. Hence, there exists some $\gamma \in \mathbb{F}_q$ such that $\gamma d_{1,1} = -\mathbf{E}_1 \mathbf{\beta}^T$. Therefore,

$$\gamma \mathbf{G}[1] = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \hline -\mathbf{E}_1 \mathbf{\beta}^T \end{pmatrix}, \quad (10)$$

From (7), (9), and (10) we have

$$\sum_{j \in L} \alpha_j \mathbf{G}[j] + \sum_{j \in J} \beta_j \mathbf{G}[j] + \gamma \mathbf{G}[1] = \mathbf{0}. \quad (11)$$

Note that the left-hand side of (11) is a linear combination of at most $2p + 1$ ($\leq k$) columns of \mathbf{G} . We argue earlier that $\beta_j \neq 0$ for every $j \in J$ and that $L \neq J$. Let $j_0 \in J \setminus L$. Then, the coefficient of $\mathbf{G}[j_0]$ in the linear combination of (11) is $\beta_{j_0} \neq 0$. Hence this linear combination is nontrivial. Therefore, we obtain a contradiction, since \mathbf{G} generates an $[n, k]$ MDS code and at the same time, has some set of at most k columns that are linearly dependent. ■

We summarize the results established so far on perfectly p -decodable μ -secure codes in the following theorem.

Theorem 16. *The Main Construction produces a coding matrix that generates a perfectly p -decodable μ -secure coding scheme when $p = 1$ for every $\mu \geq 0$, $\mu < k$, and when $\mu = 0$, 1 for every $p \geq 1$, $p|(k - \mu)$.*

D. Existence of Perfect Coding Schemes Over Large Fields

For general p and μ , the Main Construction may not produce a perfectly p -decodable μ -secure coding matrix. In order to achieve a perfect code, there is another property that the superregular matrix \mathbf{G}'' (used in Step 1 of the Main Construction) must satisfy. We first define this property in Definition 17 and then proceed to prove that there always exists a superregular matrix satisfying this property as long as the field size is sufficiently large.

Definition 17. A $k \times n$ matrix \mathbf{G}'' ($k \leq n$) is said to be (p, μ) -superregular if it is superregular and also satisfies the following additional property. Let \mathbf{G}'' be written in the block form

$$\mathbf{G}'' = \left(\begin{array}{c|c|c} \mathbf{A}'' & \mathbf{B}'' & \mathbf{C}'' \\ \hline \mathbf{D} & \mathbf{E} & \mathbf{F} \end{array} \right)_{\mu}^{k-\mu, n-k}. \quad (12)$$

Let \mathbf{B}_i'' be the i th $p \times p$ submatrix lying on the main diagonal of \mathbf{B}'' . Note that the entries outside these $p \times p$ block matrices are nonzeros.

$$\mathbf{B}'' = \begin{pmatrix} \mathbf{B}_1'' & & & \\ & \mathbf{B}_2'' & & \\ & & \ddots & \\ & & & \mathbf{B}_{\frac{k-\mu}{p}}'' \end{pmatrix}.$$

Let \mathbf{E}_i be the i th $\mu \times p$ submatrix of \mathbf{E} such that $\mathbf{E} = (\mathbf{E}_1 | \mathbf{E}_2 | \dots | \mathbf{E}_{\frac{k-\mu}{p}})$. For each $i \in [(k - \mu)/p]$ we consider the block matrix \mathbf{H}_i given below.

$$\mathbf{H}_i = \left(\begin{array}{c|c} \mathbf{O} & \mathbf{B}_i'' \\ \hline \mathbf{D} & \mathbf{E}_i \end{array} \right)_{\mu}^p. \quad (13)$$

Then it is required that for every $i \in [(k - \mu)/p]$, deleting simultaneously an arbitrary row among the first p rows of \mathbf{H}_i and an arbitrary column among the first μ columns of \mathbf{H}_i always results in an invertible $(\mu + p - 1) \times (\mu + p - 1)$ matrix.

Let $\mathbf{\Xi}_{\mathbf{G}''} = (\xi_{i,j})$ be a $k \times n$ matrix where $\xi_{i,j}$'s are indeterminates over some \mathbb{F}_q . The subscript \mathbf{G}'' simply means that at a certain time, we will replace the entries $\xi_{i,j}$'s of $\mathbf{\Xi}_{\mathbf{G}''}$ by appropriate elements of \mathbb{F}_q to obtain a matrix \mathbf{G}'' over \mathbb{F}_q , which is to be used in the Main Construction. We represent $\mathbf{\Xi}_{\mathbf{G}''}$ as a block matrix as follows.

$$\mathbf{\Xi}_{\mathbf{G}''} = \left(\begin{array}{c|c|c} \mathbf{\Xi}_{\mathbf{A}''} & \mathbf{\Xi}_{\mathbf{B}''} & \mathbf{\Xi}_{\mathbf{C}''} \\ \hline \mathbf{\Xi}_{\mathbf{D}} & \mathbf{\Xi}_{\mathbf{E}} & \mathbf{\Xi}_{\mathbf{F}} \end{array} \right)_{\mu}^{k-\mu, n-k}. \quad (14)$$

Let $f_{\text{super}}(\dots, \xi_{i,j}, \dots) \in \mathbb{F}_q[\dots, \xi_{i,j}, \dots]$ be the product of the determinants of all square submatrices of $\mathbf{\Xi}_{\mathbf{G}''}$.

Let $\mathbf{\Xi}_{B_i''}$ be the i th $p \times p$ submatrix lying on the main diagonal of $\mathbf{\Xi}_{\mathbf{B}''}$.

$$\mathbf{\Xi}_{\mathbf{B}''} = \begin{pmatrix} \mathbf{\Xi}_{B_1''} & & & \\ & \mathbf{\Xi}_{B_2''} & & \\ & & \ddots & \\ & & & \mathbf{\Xi}_{B_{\frac{k-\mu}{p}}''} \end{pmatrix}. \quad (15)$$

Let $\mathbf{\Xi}_{\mathbf{E}}$ be the i th $\mu \times p$ submatrix of $\mathbf{\Xi}_{\mathbf{E}}$ such that $\mathbf{\Xi}_{\mathbf{E}} = (\mathbf{\Xi}_{\mathbf{E}_1} | \mathbf{\Xi}_{\mathbf{E}_2} | \dots | \mathbf{\Xi}_{\mathbf{E}_{(k-\mu)/p}})$. For each $i \in [(k - \mu)/p]$ we consider the block matrix $\mathbf{\Xi}_{H_i}$ given below.

$$\mathbf{\Xi}_{H_i} = \left(\begin{array}{c|c} \mathbf{O} & \mathbf{\Xi}_{B_i''} \\ \hline \mathbf{\Xi}_{\mathbf{D}} & \mathbf{\Xi}_{\mathbf{E}_i} \end{array} \right)_{\mu}^p. \quad (16)$$

Let $f_i(\dots, \xi_{i,j}, \dots) \in \mathbb{F}_q[\dots, \xi_{i,j}, \dots]$ be the product of determinants of all square $(\mu + p - 1) \times (\mu + p - 1)$ submatrices of $\mathbf{\Xi}_{H_i}$ obtained by deleting simultaneously one arbitrary row among the first p rows and one arbitrary column among the first μ columns. Let $f_{(\mu,p)} = \prod_{i=1}^{\frac{k-\mu}{p}} f_i \in \mathbb{F}_q[\dots, \xi_{i,j}, \dots]$.

Lemma 18. *The polynomials f_{super} and $f_{(\mu,p)}$ defined as above both are not identically zero.*

Proof: In this proof we use the standard definition of determinant to calculate a determinant of a matrix. More specifically, if $\mathbf{P} = (p_{i,j})$ is any square matrix of order r then the determinant

$$\det(\mathbf{P}) = \sum_{\sigma \in \mathcal{S}_r} \text{sgn}(\sigma) \prod_{i=1}^r p_{i,\sigma(i)}, \quad (17)$$

where \mathcal{S}_r denotes the symmetric (permutation) group on r elements, and $\text{sgn}(\sigma)$ denotes the sign of the permutation σ .

We first prove that f_{super} is not identically zero. It suffices to show that the determinant of each square submatrix of $\Xi_{\mathbf{G}''}$ is not identically zero. According to (17), such a determinant is a sum of $(\mu + p - 1)!$ monomials (with appropriate signs ± 1) of $\xi_{i,j}$'s. Those monomials are obviously pairwise distinct. Therefore, their sum is not identically zero.

We now show that $f_{(\mu,p)}$ is not identically zero. It suffices to show that for each $i \in [(k - \mu)/p]$, deleting simultaneously a row among the first p rows and a column among the first μ columns of Ξ_{H_i} (given in (16)) results in a matrix with not identically zero determinant. For convenience, let us refer to such a matrix as \mathbf{P} . Then \mathbf{P} is a square matrix of order $(\mu + p - 1)$. All entries of \mathbf{P} are indeterminates $\xi_{i,j}$, except for those lying in the $(p - 1) \times (\mu - 1)$ top-left submatrix.

$$\mathbf{P} = \left(\begin{array}{c|c} \mathbf{O} & (\xi_{i,j})_{\mu}^p \\ \hline (\xi_{i,j})_{\mu} & (\xi_{i,j})_{\mu}^{p-1} \end{array} \right)^{\mu+p-1}. \quad (18)$$

Therefore, all entries on the antidiagonal (the diagonal goes from the lower-left corner to the upper-right corner of the matrix) are nonzero. These $\mu + p - 1$ antidiagonal entries form a unique nonzero monomial of $\xi_{i,j}$'s in the formula for determinant of \mathbf{P} in (17), which cannot be canceled out by any other monomial. Hence, $\det(\mathbf{P})$ is not identically zero as a polynomial. ■

Lemma 19. *For every $0 \leq \mu < k$ and $1 \leq p \leq k - \mu$, $p|(k - \mu)$, there always exists a (p, μ) -superregular $k \times n$ matrix \mathbf{G}'' over any sufficiently large field.*

Proof: According to Lemma 18, the polynomial $f_{\text{super}} f_{(\mu,p)} \in \mathbb{F}_q[\dots, \xi_{i,j}, \dots]$ is not identically zero. By [17, Lemma 4], for sufficiently large q , there exist $g''_{i,j} \in \mathbb{F}_q$ such that $f_{\text{super}}(\dots, g''_{i,j}, \dots) \neq 0$ and $f_{(\mu,p)}(\dots, g''_{i,j}, \dots) \neq 0$. The former condition guarantees that the $k \times n$ matrix $\mathbf{G}'' = (g''_{i,j})$ is a superregular matrix, while the latter further implies that \mathbf{G}'' is (p, μ) -superregular, according to Definition 17. Thus, a (p, μ) -superregular $k \times n$ matrix \mathbf{G}'' always exists over any sufficiently large field. ■

Lemma 20. *Suppose that the matrix \mathbf{G}'' used in Step 1 of the Main Construction is (p, μ) -superregular and that \mathbf{G} is the*

resulting coding matrix.

$$\mathbf{G} = \left(\begin{array}{c|ccc|c} & \mathbf{B}_1'' & & & & \\ & & \mathbf{B}_2'' & & & \\ & & & \ddots & & \\ & & & & \mathbf{B}_{(k-1)/p}'' & \\ \hline \mathbf{D} & \mathbf{E}_1 & \mathbf{E}_2 & \cdots & \mathbf{E}_{(k-1)/p} & \mathbf{F} \end{array} \right), \quad (19)$$

where each \mathbf{E}_i is a $\mu \times p$ submatrix of \mathbf{E} such that $\mathbf{E} = (\mathbf{E}_1 | \mathbf{E}_2 | \cdots | \mathbf{E}_{(k-1)/p})$. For each $i \in [(k - \mu)/p]$ let

$$\mathbf{H}_i = \left(\begin{array}{c|c} \mathbf{O} & \mathbf{B}_i'' \\ \hline \mathbf{D} & \mathbf{E}_i \end{array} \right)_{\mu}^p. \quad (20)$$

Then any set of $\mu + p - 1$ columns of \mathbf{H}_i that consists of the last p columns and some $\mu - 1$ columns among the first μ columns generates a $[\mu + p, \mu + p - 1, 2]$ MDS code. Thus, such a set of $\mu + p - 1$ columns of \mathbf{H}_i , as well as the corresponding set of $\mu + p - 1$ columns of \mathbf{G} , never generates a nontrivial vector of weight less than two.

Proof: Note that since \mathbf{G}'' is (p, μ) -superregular, deleting simultaneously an arbitrary row among the first p rows and an arbitrary column among the first μ columns of \mathbf{H}_i always results in an invertible matrix of order $\mu + p - 1$.

Let \mathbf{K} be the submatrix of \mathbf{H}_i obtained by deleting some column among the first μ columns of \mathbf{H}_i . In order to show that the columns of \mathbf{K} generate an MDS code, it suffices to prove that every square submatrix \mathbf{Q} of order $\mu + p - 1$ of \mathbf{K} is invertible. Indeed, if \mathbf{Q} is obtained by deleting one row among the first p rows of \mathbf{K} , then it is obviously invertible, due to the property of \mathbf{H}_i specified earlier. Suppose that \mathbf{Q} is obtained by deleting one row among the last μ rows of \mathbf{K} . Then $\det(\mathbf{Q})$ is equal to the product of $\det(\mathbf{B}_i'')$ and the determinant of a square submatrix of order $\mu - 1$ of \mathbf{D} . As both \mathbf{B}_i'' and \mathbf{D} are superregular, we conclude that the determinant of \mathbf{Q} must be nonzero. Therefore, \mathbf{Q} is invertible. ■

Theorem 21. *For every $0 \leq \mu < k$ and $1 \leq p \leq k - \mu$, $p|(k - \mu)$, such that $k \geq 2(\mu + p) - 1$, there exists a perfectly p -decodable μ -secure coding scheme over any sufficiently large field \mathbb{F}_q .*

Proof: In Step 1 of the Main Construction let us choose \mathbf{G}'' to be a (p, μ) -superregular matrix over a sufficiently large field \mathbb{F}_q . Let \mathbf{G} be the coding matrix produced by the Main Construction. We present \mathbf{G} in the block form as shown in the Main Construction.

$$\mathbf{G} = \left(\begin{array}{c|ccc|c} & \mathbf{B}_1'' & & & & \\ & & \mathbf{B}_2'' & & & \\ & & & \ddots & & \\ & & & & \mathbf{B}_{(k-1)/p}'' & \\ \hline \mathbf{D} & \mathbf{E}_1 & \mathbf{E}_2 & \cdots & \mathbf{E}_{(k-1)/p} & \mathbf{F} \end{array} \right). \quad (21)$$

Note that each \mathbf{E}_i is a $\mu \times p$ submatrix of \mathbf{E} such that $\mathbf{E} = (\mathbf{E}_1 | \mathbf{E}_2 | \cdots | \mathbf{E}_{(k-1)/p})$.

According to Lemma 10 and Lemma 11, \mathbf{G} satisfies (P1) and (P2). It remains to show that \mathbf{G} satisfies (P3) - weak security.

According to Lemma 3, we aim to show that no subsets of $(\mu + p - 1)$ columns of \mathbf{G} generate the unit vector \mathbf{e}_i for every $i \in [k - \mu]$. Without loss of generality, we can assume, by contradiction, that a subset L of some $(\mu + p - 1)$ columns of \mathbf{G} generates the unit vector \mathbf{e}_1 . For simplicity, we slightly abuse the notation and also use L to denote the set of indices of the columns in L . Then there exist some coefficients $\alpha_j \in \mathbb{F}_q$ ($j \in L$) such that

$$\sum_{j \in L} \alpha_j \mathbf{G}[j] = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad (22)$$

where $\mathbf{G}[j]$ is the j th column of \mathbf{G} . For convenience, we assign $\alpha_j = 0$ for all $j \in [n] \setminus L$. Let $J = \{\mu + 1, \mu + 2, \dots, \mu + p\}$.

As \mathbf{B}_1'' is invertible, there exist some coefficients $\beta_j \in \mathbb{F}_q$ such that

$$\sum_{j \in J} \beta_j \mathbf{G}[j] = \begin{pmatrix} -1 \\ 0 \\ \vdots \\ 0 \\ \mathbf{E}_1 \boldsymbol{\beta}^T \end{pmatrix}, \quad (23)$$

where $\boldsymbol{\beta} = (\beta_{\mu+1}, \dots, \beta_{\mu+p})$. According to Lemma 8, any set of less than p columns of \mathbf{B}_1'' would never generate a nontrivial vector of weight less than two. Therefore, from (23), we deduce that $\beta_j \neq 0$ for every $j \in J$. Moreover, as \mathbf{D} is invertible, there exist some coefficients $\gamma_j \in \mathbb{F}_q$ such that

$$\sum_{j \in [\mu]} \gamma_j \mathbf{D}[j] = -\mathbf{E}_1 \boldsymbol{\beta}^T,$$

where $\mathbf{D}[j]$ is the j th column of \mathbf{D} . Therefore,

$$\sum_{j \in [\mu]} \gamma_j \mathbf{G}[j] = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ -\mathbf{E}_1 \boldsymbol{\beta}^T \end{pmatrix}. \quad (24)$$

From (22), (23), and (24) we have

$$\sum_{j \in L} \alpha_j \mathbf{G}[j] + \sum_{j \in J} \beta_j \mathbf{G}[j] + \sum_{j \in [\mu]} \gamma_j \mathbf{G}[j] = \mathbf{0}. \quad (25)$$

Note that the left-hand side of (25) is a linear combination of at most $2(\mu + p) - 1$ ($\leq k$) columns of \mathbf{G} . We consider the following three cases and aim to obtain contradictions in all cases.

Case 1. $\exists j_0 \in J : \alpha_{j_0} = 0$.

We argue earlier that $\beta_j \neq 0$ for every $j \in J$. Moreover, in this case, there exists $j_0 \in J$ such that $\alpha_{j_0} = 0$. Hence,

the linear combination in (25) must be nontrivial, since at least one vector, namely $\mathbf{G}[j_0]$, has a nonzero coefficient β_{j_0} . Therefore, we obtain a contradiction, since \mathbf{G} generates an $[n, k]$ MDS code and at the same time, has some set of at most k columns that are linearly dependent.

Case 2. $\forall j \in J : \alpha_j \neq 0$ and $L \setminus J \subseteq [\mu]$.

Note that in this case, L consists of p columns indexed by J and some $\mu - 1$ of the first μ columns of \mathbf{G} . According to Lemma 20, L does not generate any nontrivial vector of weight less than two. This conclusion contradicts (22).

Case 3. $\forall j \in J : \alpha_j \neq 0$ and $L \setminus J \not\subseteq [\mu]$.

In this case, L consists of p columns indexed by J , at most $\mu - 1$ of the first μ columns of \mathbf{G} , and probably some columns indexed by elements in the set $[n] \setminus ([\mu] \cup J)$. If $\alpha_j = 0$ for all $j \in L \setminus ([\mu] \cup J)$, then similar to Case 2, we obtain a contradiction due to Lemma 20 and (22).

If there exists $j \in L \setminus ([\mu] \cup J)$ such that $\alpha_j \neq 0$, then (25) presents a nontrivial linear combination of at most k columns of \mathbf{G} , which is a zero vector. This assertion contradicts the fact that as \mathbf{G} generates an $[n, k]$ MDS code, any set of k columns of \mathbf{G} must be linearly independent. ■

V. CONCLUSION

We propose in this work a method to construct erasure coding schemes which are not only (strongly and weakly) secure but also partially decodable. The partial decodability feature is extremely important in applications such as media streaming, where it is usually not necessarily for the user to download the whole file before he or she can start the playback.

The type of erasure coding scheme developed in our work offer the flexibility between security and partial decodability. The system designer can adjust the security parameter and the partial decodability parameter accordingly to obtain any possible mixture of security and exposure of the stored data. We emphasize that we can construct an erasure code which is both secure and partially decodable without adding any extra storage overhead compared to a merely secure erasure code studied in the literature.

REFERENCES

- [1] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Amsterdam: North-Holland, 1977.
- [2] H. Weatherspoon and J. Kubiatowicz, "Erasure coding vs. replication: A quantitative comparison," in *Revised Papers from the First International Workshop on Peer-to-Peer Systems*, ser. IPTPS '01, 2002, pp. 328–338.
- [3] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin, "Erasure coding in windows azure storage," in *Proc. USENIX Conf. Annual Technical Conference (ATC)*, 2012, pp. 2–2.
- [4] A. Thusoo, Z. Shao, S. Anthony, D. Borthakur, N. Jain, J. S. Sarma, R. Murthy, and H. Liu, "Data warehousing and analytics infrastructure at Facebook," in *Proc. ACM SIGMOD Int. Conf. on Management of Data (SIGMOD)*, 2010, pp. 1013–1020.
- [5] D. Ford, F. Labelle, F. I. Popovici, M. Stokely, V.-A. Truong, L. Barroso, C. Grimes, and S. Quinlan, "Availability in globally distributed storage systems," in *Proc. USENIX Conf. Operating Syst. Design Implementation*, 2010, pp. 1013–1020.

- [6] L. H. Ozarow and A. D. Wyner, "The wire-tap channel II," *Bell Syst. Tech. J.*, vol. 63, pp. 2135–2157, 1984.
- [7] A. Subramanian and S. W. McLaughlin, "MDS codes on the erasure-eraser wiretap channel," 2009. [Online]. Available: <http://arxiv.org/abs/0902.3286>
- [8] H. Yamamoto, "Secret sharing system using (k, L, n) threshold scheme," *Electronics and Communications in Japan (Part I: Communications)*, vol. 69, no. 9, pp. 46–54, 1986.
- [9] M. Iwamoto and H. Yamamoto, "Strongly secure ramp secret sharing schemes for general access structures," *Inf. Process. Lett.*, vol. 97, no. 2, pp. 52–57, 2006.
- [10] P. F. Oliveira, L. Lima, T. T. V. Vinhoza, J. Barros, and M. Médard, "Coding for trusted storage in untrusted networks," *IEEE Trans. Inform. Forensics Security*, vol. 7, no. 6, pp. 1890–1899, 2012.
- [11] S. H. Dau, W. Song, and C. Yuen, "On block security of regenerating codes at the MBR point for distributed storage systems," in *Proc. Int. IEEE Symp. Inform. Theory (ISIT)*, 2014, pp. 1967–1971.
- [12] S. Kadhe and A. Sprintson, "Weakly secure regenerating codes for distributed storage," in *Int. Symp. Network Coding (NetCod)*, 2014, pp. 1–6.
- [13] N. Cai and T. H. Chan, "Theory of secure network coding," *Proceedings of the IEEE*, vol. 99, no. 3, pp. 421–437, 2011.
- [14] K. Bhattad and K. R. Narayanan, "Weakly secure network coding," in *Proc. 1st Workshop on Network Coding, Theory, and Application (NetCod)*, 2005.
- [15] R. M. Roth and A. Lempel, "On MDS codes via Cauchy matrices," *IEEE Trans. Inform. Theory*, vol. 35, no. 6, pp. 1314–1319, 1989.
- [16] J. Lacan and J. Fimes, "Systematic MDS erasure codes based on Vandermonde matrices," *IEEE Commun. Lett.*, vol. 8, no. 9, pp. 570–572, 2004.
- [17] T. Ho, M. Médard, R. Koetter, D. R. Karger, M. Effros, J. Shi, and B. Leong, "A random linear network coding approach to multicast," *IEEE Trans. Inform. Theory*, vol. 52, no. 10, pp. 4413–4430, 2006.

APPENDIX

A. Proof of Lemma 3

The *only if* direction is obvious, according to the discussion below Lemma 3. It remains to prove the *if* direction.

Let \tilde{s} and \tilde{r} be the guessed value (by the adversary) for the real file s and for the random vector r , respectively. By accessing m storage nodes, the adversary obtains some linear combinations of the file symbols s_i 's and the random symbols r_j 's, namely $a = (s \mid r)M$. We aim to show that for every $i \in [k - \mu]$, all possible values for s_i are equally probable.

Choose an arbitrary $i \in [k - \mu]$ and an arbitrary element $\tilde{v} \in \mathbb{F}_q$. It suffices to show that the system of linear equations

$$\begin{cases} (\tilde{s} \mid \tilde{r})M = a \\ (\tilde{s} \mid \tilde{r})e_i^T = \tilde{v} \end{cases} \quad (26)$$

always has the same number of solutions $(\tilde{s} \mid \tilde{r})$ for every choice of $\tilde{v} \in \mathbb{F}_q$. Here e_i denotes the unit vector with a one at the i th coordinate and zeroes elsewhere. It is a basic fact from linear algebra that the solution set for the system (26) above, if nonempty, is an affine space, which is the sum of one solution of (26) and the solution space of the corresponding homogeneous system. Therefore, if the system (26) always has at least one solution for every \tilde{v} , then it would have the same number of solutions for every \tilde{v} . Therefore, it remains to prove that this system always has a solution for every choice of $\tilde{v} \in \mathbb{F}_q$.

Note that we have

$$\begin{cases} (s \mid r)M = a \\ (s \mid r)e_i^T = s_i \end{cases} \quad (27)$$

By subtracting the corresponding equations in the two systems (26) and (27) and let $x = (\tilde{s} - s \mid \tilde{r} - r)$ be the new unknowns, we obtain the following system

$$\begin{cases} xM = 0 \\ xe_i^T = \tilde{v} - s_i \end{cases} \quad (28)$$

It is clear that the system (26) has a solution if and only if the system (28) has a solution. Hence, it suffices to show that the system (28) always has a solution for every choice of $\tilde{v} \in \mathbb{F}_q$.

We claim that there exists some $\bar{x} \in \mathbb{F}_q^k$ satisfying $\bar{x}M = 0$ and $\bar{x}e_i^T \neq 0$. Then it is obvious that

$$x^* = \frac{\tilde{v} - s_i}{\bar{x}e_i^T} \bar{x}$$

would be a solution of (28), and hence the proof follows. Indeed, if $\bar{x}e_i^T = 0$ for every \bar{x} satisfying $\bar{x}M = 0$ then e_i^T must belong to the orthogonal complement of the solution space of the system $xM = 0$, which is precisely the column space of M . However, according to our assumption, the column space of M does not contain e_i . Thus, there must exist some $\bar{x} \in \mathbb{F}_q^k$ satisfying $\bar{x}M = 0$ and $\bar{x}e_i^T \neq 0$, as claimed above.